DOCUMENT RESUME

ED 227 141 TM 830 141

AUTHOR Huberty, Carl J; Smith, Janet C.
TITLE Validation in Group Membership Prediction.
PUB DATE Aug 82
NOTE 36p.; Paper presented at the Annual Meeting of the American Psychological Association (Washington, DC, August 23-28, 1982).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Classification; *Cluster Grouping; *Discriminant Analysis; Estimation (Mathematics); Multidimensional Scaling; *Predictive Measurement; Regression (Statistics); *Validity
IDENTIFIERS Multiple Measures Approach

ABSTRACT

Predictive discriminant analysis involves a technique used in multivariate classification, i.e., in predicting membership in well-defined groups for units on which multiple measures are available. The validation (assessment) of group membership predictions pertains to two problems: estimating true proportions of correct classifications (i.e., hit rates) and determining, on the basis of one data set, a classification rule to use with other data sets. Two approaches to the hit rate problem are the shrinkage formula and external analysis. The formula approach has been developed only for the two-group case. Of the two external analysis methods, holdout and leave-one-out, the latter is preferred. The form of a classification rule considered is a set of k linear classification functions (LCF's). The problem of determining a classification rule has two dimensions: determining LCF weight estimates, and determining which predictors to delete from consideration in the final formulation of the rule. Jackknife methodology is used to obtain LCF weight estimates and to assess the variability of the estimates. This assessment leads to approximate t statistics which may be used to determine which of the predictors, if any, to delete. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED227141

Validation in Group Membership Prediction

Carl J Huberty Janet C. Smith

University of Georgia

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Paper presented in the symposium, "The Many Faces of Cross-Validation," at the Annual Meeting of the American Psychological Association, Washington, August 1982.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C. J. Huberty

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 830 141

Validation in Group Membership Prediction

Introduction 1

Hit Rate Estimation 3

- Formula Approach 4
- External Analysis Approach 5
  - Holdout Method 6
  - Leave-One-Out Method 7
- Sample Size 8
- Some Illustrations 8

LCF Weight Estimation 10

- Jackknifed Estimators 11
- An Illustration 13

Predictor Deletion 14

Summary 17

Appendix 20

Abstract

Two aspects of "validation" in predictive discriminant analysis are considered. One aspect is the estimation of classification hit rates. Two estimation approaches are reviewed: shrinkage formula, and external analysis. The snrinkage formulas leave something to be desired. Of the two external analysis methods, holdout and leave-one-out, the latter is generally preferred The second aspect is the determination of a good classification rule, which includes the estimation of linear classification function weights and the deletion of predictors. These two problems are simultaneously attacked using jackknife methodology.

Validation in Group Membership Prediction

Introduction

In situations calling for multiple linear regression, prediction rules are often formulated by utilizing the (ordinary) least squares criterion. By so doing, optimality in some senses is achieved. The sample regression weights are optimum in the sense that they are unbiased, maximum-likelihood estimates (if the usually postulated linear model is in fact correct). The weights are also optimum in that the correlation between scores on the resulting linear composite of the predictors with scores on the criterion is maximized. That is, no other linear composite will yield a larger correlation for the data on hand. The weights may not be optimal, however, for subsequent samples.

In a regression context, the multiple correlation coefficient, R, is typically used as an index of accuracy of prediction for the set of predictors being studied. As a descriptive index, R may be considered appropriate; for inferential purposes, however, R is deficient. It is well known (see, e.g., Ruberty & Mourad, 1980) that R is a biased estimator for the true validity coefficient, ρ_v [$E(R) > \rho_v$].

In sum, then, there are two validity concerns in multiple regression analysis: (1) assessing the accuracy of a prediction rule; and (2) determining a good prediction rule to use with new data. A dimension of the second concern which needs to be considered is that of determining those predictors to be deleted in formulating the final rule.

There are somewhat similar inference concerns in the context of multivariate group membership prediction, one aspect of "discriminant analysis."

To attack this prediction problem one can use a collection of techniques and methods known as predictive discriminant analysis (PDA). One purpose of using PDA is to determine a prediction, or classification, rule for assigning experimental units, on the basis of a set of p response measures, into one of k ($\geq$ 2) criterion groups. Associated prototypical research questions are: (a) To what extent can degree completion (graduate versus non-graduate) be predicted for master level aspirants?, and (b) Which of four typical bachelor graduation major-field profiles is most closely resembled by the profile of an incoming college student? Information acquired in studying the first question may be of practical importance for admissions committees; the second question may be of more "theoretical" importance, although of some practical importance for academic counselors.

[For the sake of completeness, the second aspect of discriminant analysis pertains to a collection of techniques called descriptive discriminant analysis (DDA) that is useful in the study of effects indicated by a multivariate analysis of variance. These techniques deal with determining the relative contribution of response variables to resultant effects, selecting variable subsets to retain for interpretation, and characterizing the structure underlying the effects.]

There are two basic questions that need to be addressed when conducting a predictive discriminant analysis: (1) How accurately can experimental units be classified into the criterion groups?, and (2) What is a good classification rule to use with units other than those for which data are available for the initial analysis? The intent of this paper is to present methods of obtaining answers to these two questions that are, to some extent, "valid." [Actually, both questions involve "assessment" rather than "validation"; the current authors concur with Stone (1974, p. 111) in preferring

the former term since the latter term has a "ring of excessive confidence about it."] The accuracy question pertains to estimating probabilities of correct classification, i.e., estimating "hit rates," and will be covered first.

Hit Rate Estimation

Estimation of a true hit rate in predictive discriminant analysis is somewhat analogous to estimation of ρ_v in regression analysis. This is the predictive accuracy problem. A candidate for an accuracy index is the apparent hit rate, the proportion of experimental units in the sample correctly reclassified using the sample-based rule. The analysis yielding such a hit rate is termed an internal PDA. Apparent hit rates are calculated from hit tables, an example of which is given below.

		Predicted Group			
		1	2	3	
Actual Group	1	n_{11}	n_{12}	n_{13}	n_1
	2	n_{21}	n_{22}	n_{23}	n_2
	3	n_{31}	n_{32}	n_{33}	n_3
		$n_{.1}$	$n_{.2}$	$n_{.3}$	N

For this table, the separate-group apparent hit rates are n_{11}/n_1, n_{22}/n_2 and n_{33}/n_3, and the total-group apparent hit rate is $(n_{11} + n_{22} + n_{33})/N$.

Just as the multiple R is a positively biased estimator for ρ_v, an apparent classification hit rate is a positively biased estimate of the true hit rate (see Michaelis, 1973, p. 230; Frank, Massey, & Morrison, 1965). Reasoning behind the positive bias of an apparent hit rate is somewhat similar to that for R. A typically used group membership prediction rule, for a k-group problem, is a set of k linear composites called linear classification functions (LCFs).[1] With known parameters, weights in these

composites are determined so as to maximize proportions of correct classifications for the given data; some optimality is lost when using parameter estimates. The near-optimizing process makes the greatest possible use of any and all idiosyncrasies of the data on hand. That is, as Mosteller and Tukey (1977, p. 37) state, "optimization capitalizes on chance." Thus, from an inference viewpoint, something other than apparent hit rates need to be used as true hit rate estimates.

Formula Approach

There are two general approaches that may be used in arriving at less biased hit rate estimates. One approach is to use a "formula" to estimate the separate-group hit rates, P_g, $g = 1, \ldots, k$. If $k = 2$ and the parameters are known, values of P_1 and P_2 are simple functions of Δ^2, the Mahalanobis index of distance between the two population mean vectors. Lachenbruch and Mickey (1968) discuss formula estimators for P_g that vary with estimators for Δ^2. An estimator for Δ^2 that has some desirable characteristics is $\tilde{D}^2 = (N - p - 3)\, D^2/(N - 2)$, where D^2 is the sample counterpart of Δ^2. Using q_g as an estimator for the prior probability of belonging to population g and $\tilde{D}^2$ as an estimator for Δ^2, formula estimators suggested by Afifi and Azen (1979, p. 293) may be utilized:

$$\hat{P}_1 = 1 - F\left(\frac{K - \tilde{D}^2/2}{\tilde{D}}\right) \text{ and } \hat{P}_2 = 1 - F\left(\frac{-K - \tilde{D}^2/2}{\tilde{D}}\right), \qquad [1]$$

where F is the standard normal distribution function and $K = \ln(q_2/q_1)$. With this formulation, costs of misclassification are assumed equal. Note that if $q_1 = q_2 = 1/2$, then $\hat{P}_1 = \hat{P}_2 = 1 - F(-\tilde{D}/2)$.

McLachlan (1975) derived a "shrinkage" formula for the two-group situation that also allows for different estimates of P_1 and P_2. McLachlan's estimator for the group g hit rate is $\hat{P}_g = 1 - Q_g$, with

$$Q_g = F(-D/2) + f(-D/2)\left\{\frac{p-1}{D\,n_g} + \frac{D}{32m}\,[4(4p-1) - D^2]\right.$$

$$+ \frac{(p-1)(p-2)}{4\,D\,n_g^2} + \frac{p-1}{64m\,n_g}\,[-D^3 + 8D(2p+1) + 16/D]$$

$$+ \frac{D}{12288\,m^2}\,[3D^6 - 4D^4(24p+7) + 16D^2(48p^2 - 48p - 53)$$

$$\left. + 192\,(-8p + 15)]\right\}, \qquad [2]$$

where F is the standard normal distribution function (i.e., $F(z) = P[Z \leq z]$), f is the standard normal density function (i.e., $(z, f(z))$ is a point on the standard normal curve), p is the number of predictor variables, D is the sample Mahalanobis distance between the two group mean vectors, and $m = n_1 + n_2 - 2$. A number of analyses conducted by the current authors indicates that for most practical use, the last term in the multiplier of $f(-D/2)$ may be ignored. This leaves a more workable shrinkage formula:

$$\hat{P}_g = 1 - F(-D/2) - f(-D/2)\left\{\frac{p-1}{D\,n_g} + \frac{D}{32m}\,[4(4p-1) - D^2]\right.$$

$$\left. + \frac{(p-1)(p-2)}{4\,D\,n_g^2} + \frac{p-1}{64m\,n_g}\,[-D^3 + 8D(2p+1) + 16/D]\right\}. \qquad [2']$$

It may be noted that the amount of shrinkage is inversely related to values of N (= m + 2) and D, and directly related to the value of p.

External Analysis Approach

The second approach to obtaining less biased hit rate estimates is not restricted to two-group situations. This approach involves the use of an external PDA. Here, the classification rule is determined from one set of

units, and then used to classify another set of units. [This approach utilizes the traditional "cross-validation" idea.] Hit rates based on the "new" data are used as estimates of the true separate-group hit rates, P_g, and of the true total-group hit rate, P_T.

Holdout method. Two ways of carrying out an external PDA are considered in the current paper--both analyses involve sample splitting and may be accomplished using widely available computer program packages. One way is to use a single splitting of the available sample into two subsamples: (i) an analysis or training sample, and (ii) a test or holdout sample. A classification rule is determined using the analysis sample data, and then applied to the test sample data. [This is the so-called "holdout method."] A hit rate estimate is the proportion of the test sample units that are correctly classified (using the rule based on the analysis sample).

An analysis using the holdout method may be accomplished using SPSS's DISCRIMINANT procedure, BMD's P7M program, or SAS's DISCRIM procedure. See the Appendix for details on how to obtain holdout analyses with the three packages.

With the holdout method one needs to decide what portion of the data to include in the test sample. Statistical theory gives little guidance, and handy rules-of-thumb are yet to be established. Only in the two-group case has a suggestion been given. Some asymptotic theory developed by Schaafsma and van Vark (1979, p. 96) suggests that the ratio of the test sample size to the analysis sample size is a function of p, the number of predictors: $[1 + (p - 1)^{1/2}]^{-1}$. This leads to "the feeling that it might be reasonable to recommend" a ratio of 0.30, at least for $p \geq 10$. This, in turn, implies that the analysis sample be approximately 75% of the total sample. [For the illustrations to be presented in this paper, the analysis samples are 70% of the total sample.]

The many problems with the holdout method are pointed out by Lachenbruch and Mickey (1968, pp. 2-3). A basic problem is the requirement of "large" samples. A method that works quite well for "small" samples, and is fairly efficient even for large samples is presented next.

Leave-one-out method. A second way of carrying out an external PDA is use the leave-one-out (L-O-O) method popularized by P. A. Lachenbruch (initially proposed in his 1965 dissertation and then published in 1967). The method involves a two-step process. First, one unit is deleted, and LCFs are determined on the remaining N - 1 units. Then these LCFs are used to classify the deleted unit into one of the k criterion groups. This process is carried out N times, and the proportions of deleted units correctly classified are used as hit rate estimates.

Hit rate estimates based on the L-O-O method may be obtained using the BMDP7M program. It is accomplished by simply specifying JACK in the DISCR paragraph in the P7M control statements.

Sample Size

How large a sample is needed for valid hit rate estimation? Of course, validity is a matter of degree; representativeness assumed, the larger the sample size the greater the validity of estimation. [That is, the larger the sample the better the assessment.] Just as for the case of the ratio of test sample size to analysis sample size, only limited guidance has been given in terms of rules-of-thumb for a minimum total sample size. This guidance, too, has been given only for the special two-group situation. Based upon results of a number of data simulations, Foley (1972) found that when n/p was greater than 3, the average internal hit rate, across replications, was "reasonably close" to the true hit rate (n is the number of units in each of two groups). As concluded by Hecker and Wegener (1978, p. 727), however, this criterion is seen as being too weak.

A tentative conclusion reached by the current authors is that to validly estimate hit rates using an external analysis, the minumum number of units in the smallest group should be approximately three times the number of predictors: $\min_{g} (n_g) \geq 3p$.

Some Illustrations

The intent of this subsection is to illustrate the use of the formulas/methods discussed to this point. A real data set is utilized. There are 77 public school reading teachers employing three different types of reading instruction ($n_1 = 33$, $n_2 = 20$, $n_3 = 24$). Eight measures were available on each of the teachers; five measures of reading knowledge, and three measures of teaching background.

The data for Groups 1 and 2 were analyzed to illustrate the use of the two formula hit rate estimators, [1] and [2'], as well as to illustrate the use of external analysis methods in the k = 2 situation. The Mahalanobis distance index value is $D^2 \doteq 0.63$. The separate-group and total-group hit rate estimates are summarized in Table 1. (The values of $\hat{P}_1$ and $\hat{P}_2$ based on

---

Insert Table 1 about here

---

formula [2] are 0.53 and 0.45, respectively.) The rule employed with the holdout and leave-one-out methods is a set of two LCFs each of which incorporates unequal estimates of prior probabilities of group membership ($q_1 = 33/53$, $q_2 = 20/53$). [Using the proportional chance criterion with q_1 and q_2 (see Huberty, 1982), the expected hit rates are $n_1^2/N^2 \doteq .39$ for Group 1 and $n_2^2/N^2 \doteq .14$ for Group 2; the total-group expected hit rate is .39 + .14 = .53.]

Some comments. As mentioned at the beginning of this section, an internal hit rate estimate (the "apparent hit rate") is positively biased and should not be used for estimation purposes. It is seen here that whereas the internal total-group hit rate is significantly better than that expected by chance (.66 versus .53), the holdout and L-0-0 estimates are not. With respect to the total-group hit rates, one might be tempted to conclude that Formula [2'] overcorrects a bit, and/or that Formula [1] and the holdout and L-0-0 methods

tend to undercorrect (see Michaelis, 1973, p. 230). There is some evidence (e.g., Lachenbruch & Mickey, 1968) to indicate that L-O-O estimates are, in general, more valid than holdout estimates.

Hit rate estimates based on the internal and external analyses involving all three groups are given in Table 2. The expected hit rates are

---

Insert Table 2 about here

0.18 for Group 1, 0.07 for Group 2, 0.10 for Group 3, and 0.35 across the three groups.

Again, the internal hit rates are misleading. And again, the more valid L-O-O total-group hit rate estimate indicates a classification accuracy not much greater than what could be attained by chance (.44 versus .35), a result which is, of course, a function of this particular data set.

LCF Weight Estimation

An attack on the second question posed in the Introduction section is now presented. This question dealt with the determination of a classification rule based on one set of data that would be a good rule to use with subsequently collected data. The rule sought may be a set of k linear classification functions with weights for all of the predictors, or a set of LCFs with weights for a smaller number of predictors. [The predictor deletion problem will be touched on in the next section.] Even though sample LCF weights have some good statistical properties as estimates, they may not be the best weights to use with new data due, in part, to variation with repeated sampling. [The amount of variation is a function of the degree of collinearity among the predictors, as well as a function of the number of predictors, p, and the total sample size, N.] Thus, methods of obtaining "better" LCF weights need to be considered.

One weight estimation method that might appeal to some methodologists is that of ridge estimation. Ridge methods have garnered some attention in multiple regression analysis, but very little attention in discriminant analysis. DiPillo (1976, 1977) employed ridge methods in attempting to improve a classification rule (with respect to increasing hit rates) for the two-group situation. DiPillo did not, however, explicitly study ridge methods for LCF weight estimation purposes. Campbell (1980) used ridge methodology in estimating linear discriminant function (LDF) weights.

Another way to obtain LCF weight estimates is to use "jackknife" methodology. The jackknife technique not only yields good point estimates of true LCF weights, but is also a means of assessing their variability. This technique, in general, involves splitting the available data set into s subsets of size h each; N = sh. Much of the research on, and most of the use of, the jackknife has centered on s = N and h = 1. This special case eliminates any arbitrariness in subset formation, and "it is probably the best form of the jackknife to use in any problem" (Miller, 1974, p. 2). This type of splitting is what was used in the L-O-O hit rate estimation procedure discussed earlier in this paper; it should be noted that with this procedure no jackknifing of estimates is actually done.

For a discussion of using jackknife methodology in a canonical correlation context see Wildt, Lambert, and Durand (1982).

Jackknifed Estimators

Let a_{gi} be the estimated weight for predictor i (i=1, ..., p) in LCF g (g = 1, ..., k) based on all N units:

$$\underset{-g}{a} = \bar{\underset{-g}{X}}' S^{-1}, \qquad [3]$$

where $\bar{\underset{-g}{X}}$ is the p x 1 vector of predictor means for the gth group, and S is the p x p within-groups covariance matrix. Let c_g denote the constant for group g:

$$c_g = -1/2\ \bar{X}_g'\ S^{-1}\ \bar{X}_g + \ln q_g\ , \qquad [4]$$

where q_g is the estimator for the prior probability of belonging to group g. [See Huberty (1975, p. 567) for elaboration.]

By deleting one unit (from the total of N), say the αth, estimators of the LCF weights and constants based on the remaining N - 1 units are denoted $a_{gi}^{(-\alpha)}$ and $c_g^{(-\alpha)}$, respectively. Following the suggestion of M. H. Quenouille given in the late 1940's, the estimators

$$J_\alpha\ (a_{gi}) = N\ a_{gi} - (N - 1)\ a_{gi}^{(-\alpha)}\ ,$$

and $$[5]$$

$$J_\alpha\ (c_g) = N\ c_g - (N - 1)\ c_g^{(-\alpha)}$$

are evaluated. Finally, the estimators

$$J\ (a_{gi}) = \frac{1}{N} \sum_{\alpha=1}^{N} J_\alpha\ (a_{gi})\ ,\ \text{or}$$

$$J\ (a_{gi}) = N\ a_{gi} - (N - 1)\ \bar{a}_{gi}^{(-\alpha)}\ ,$$

and $$[6]$$

$$J\ (c_g) = N\ c_g - (N - 1)\ \bar{c}_g^{(-\alpha)}$$

are derived. The J_α values in [5] were, about a decade later, called "pseudo-values" by J. W. Tukey, who also created the name "jackknifed estimator" for expressions like those in [6].

The calculation routine for arriving at jackknifed estimates of the LCF weights and constants, then, is a three-step process:

(i) Repeatedly delete one unit at a time, each time getting an adjusted $\bar{X}_g$ vector (if the deleted unit is from group g), and an adjusted inverse, S^{-1} -- the adjusted vector and matrix inverse are easily obtained from the unadjusted vector and matrix inverse using formulations derived by M. S.

Bartlett (see Lachenbruch, 1967, pp. 641-642). This adjusted vector and matrix are used in equations [3] and [4] to get "adjusted" weights and constants.

(ii) The pseudo-values, J_α, in [5] are next obtained.

(iii) Finally, the jackknifed weights and constants are calculated using [6].

A question may have arisen in the mind of the reader: "Why go through this apparently complex process to obtain these estimates?" First of all, the calculations are fairly easy to accomplish. Secondly, the estimates so obtained may be expected to be better than those based on [3] and [4] in that jackknifed estimators are generally less biased (Miller, 1974) and less variable over repeated sampling. It is for these reasons that we think jackknife methodology has some promise as a means of arriving at a classification rule better than that determined using [3] and [4].

An Illustration

The three-group data set briefly described near the end of the last section is used to illustrate jackknife estimation of the LCFs. The internally estimated LCF weights and constants (using Equations [3] and [4]) and the jackknifed estimates are given in Table 3. A question naturally

---

Insert Table 3 about here

---

arises: Are the jackknifed LCFs "better" than the internal LCFs? Since a goal of PDA is to arrive at a set of LCFs to use with "new" data, this question can best be answered by comparing the two sets of LCFs on some new data. The availability of some additional data enabled us to make a "sneak preview." The new data set consisted of a total of 75 teachers with 32, 20, and 23, in the respective groups.

The total-group hit rate based on the internal LCFs applied to the new data set is 0.36; that based on the jackknifed LCFs is 0.40. This reflects only a difference of three correct classifications. Even though the hit rate improved from one below that expected by chance (0.38) to one above that expected by chance, the improvement is not that substantial. A real test, however, would involve the application of the jackknifed LCFs to additional new data sets (presumably based on samples from a common population).

It is conjectured that some peculiarities of the data sets used here may have precluded a valid assessment of the goodness of jackknifed LCFs. The three groups are not clearly separated, as evidenced by the estimated total-group hit rate of less than 0.45. That is, either group membership (as defined with these groups) cannot be very accurately predicted, or better predictor variables were excluded from study. At any rate, the current authors have some confidence in jackknifed LCFs. Of course, more extended empirical comparisons are needed.

Predictor Deletion

It is clear from examing the two-group shrinkage formula, [2], derived by G.J. McLachlan, that shrinkage in estimated predictive accuracy is directly related to the number of predictors. Furthermore, given a set of p predictors, a subset of less than p may, in fact, yield more accurate prediction than that yielded by the total set (see, e.g., Hecker & Wegener, 1978; Murray, 1977). Thus, it behooves the researcher to determine if a good rule based on p predictors can be improved upon by deleting one or more predictors. [From a practical viewpoint, it is desirable to delete predictors, if possible, so that subsequent data collection expense would be reduced.]

The predictor deletion problem in predictive discriminant analysis (PDA) has not been investigated to any great extent in the literature. The

various approaches to "variable selection" in discriminant analysis deal almost exclusively with descriptive discriminant analysis (DDA) situations which are not considered appropriate in PDA situations. In DDA, criteria for deleting variables pertain to group separation (e.g., Wilks' lambda) but not to classification accuracy. [See McKay & Campbell, 1982; Huberty, 1982; McLachlan, 1980; Habbema & Hermans, 1977.]

One possible approach to the deletion problem is to consider discarding predictors whose LCF weights are relatively unstable, i.e., highly variable over repeated sampling. Exact distribution theory regarding LCF weights has not yet been developed. The variability of the weights can be assessed, however, using jackknife methodology (not assuming any particular distribution form of the predictors). The standard deviation of the pseudo-values, $J_\alpha(a_{gi})$ in [5], divided by $\sqrt{N}$, can be used as an estimate of the standard error of a_{gi}. The quotient of the jackknifed weight, $J(a_{gi})$, and the estimated standard error of a_{gi}, $s_{a_{gi}}$, may be considered an approximate Student t statistic $[J(a_{gi}) / s_{a_{gi}}]$. As suggested by Mosteller and Tukey (1977, p. 136), the associated degrees-of-freedom value would be N-1. These "distribution-free" t-statistics might be useful in deleting predictors.

To illustrate predictor deletion and its effect on classification accuracy, consider the data set employed in the previous sections. The t-ratios for the three-group analysis are reported in Table 4. If one were

---

Insert Table 4 about here

---

to stick by his/her (strict) statistical guns ($t_{.95}(76) = 1.67$), it might be concluded that only the weight for Predictor 3 should be considered "seriously different from zero." Being more liberal, it was decided to investigate the

effect of deleting three rather than seven predictors. On the basis of t-ratios for a given predictor on all three LCFs, Predictors 1, 4, and 6 were deleted. Three questions then emerge: (1) What is a good classification rule based on the retained subset of the original set of predictors?; (2) What effect does the deletion of predictors have on classification accuracy?; and (3) Does another subset of the same size yield a better rule?

As argued in the immediately preceding section, a good rule is one defined by a set of k LCFs with jackknifed weights (and constants). The weights and associated t-ratios for the five retained predictors are given in Table 5.

Insert Table 5 about here

The effect of using five predictors versus eight predictors may be investigated in two ways. First, the L-O-O total-group hit rate using all eight predictors and the 77 teachers is, as reported in Table 2, 0.44. The total-group L-O-O hit rate estimate using the five retained predictors is 0.40. Hence a drop of only about 4% in classification accuracy is realized when using the 77 teachers. On this basis alone, it appears that Predictors 1, 4, and 6 could be deleted with no appreciable loss in classification accuracy. A second assessment may be made by applying the two rules--one based on eight predictors and one based on five predictors--on a new data set. Applying the two rules (with jackknifed weights and constants) to the new data on 75 teachers as was used in the immediately preceding section yielded, again, comparable classification accuracy, 0.40 with eight predictors versus 0.37 with five predictors--a difference of two out of 75 correct classifications.

Now for the third question. As mentioned at the outset of this section, most discriminant analysis variable selection approaches that have been

proposed involve a criterion of group mean differences. A popular means of determining a subset of a given size to retain is to use a forward (stepwise) selection computer program such as BMDP7M. Applied to our original data set on 77 teachers, the five "best" predictors so selected are 1, 2, 3, 7, and 8. The total-group L-0-0 hit rate using these five predictors on the original set of 77 teachers is 0.43. Applying the jackknifed weights for Predictors 1, 2, 3, 7, and 8 to our new data set of 75 teachers, however, yielded a total-group hit rate of 0.33. This is clearly no better than the hit rate of 0.37 obtained using the five predictors (2, 3, 5, 7, and 8) determined by the t-ratios for the jackknifed weights.

Of course, the above paragraph does not provide anything close to a final answer to the third question. A bullish approach to arriving at a final answer would involve comparisons with the other 54 subsets of size 5. Deferring to a less laborious attack, one might consider determining the five most "important" predictors and use these five as a final subset. This might be accomplished as follows. Using the total-group L-0-0 hit rate as a criterion of importance, the hit rate based on all eight predictors is found first. [In a given situation, a particular separate-group hit rate may be used as the criterion of importance.] Then each predictor is deleted in turn to determine hit rates based on the remaining seven. The most important predictor is that which when deleted causes the largest drop in the hit rate (from the hit rate based on all eight predictors). The second most important predictor is that which when deleted causes the next largest drop. And so on to the eighth or least important predictor. These eight analyses indicated the following predictor ordering: 2, 1, 8, 5 and 6 and 7, 3 and 4. Thus another "best" subset of size 5 consists of Predictors 2, 1, 8, and two of 5, 6, 7. The total-group L-0-0 hit rate using Predictors 2, 1, 8, 5, 7

on the original set of 77 teachers is 0.45. [The other two subsets of size five had the same hit rate.] Applying the jackknifed weights for these predictors to the new set of 75 teachers yielded a total-group hit rate of 0.32. Again, this is clearly no better than the hit rate of 0.37 obtained using the five predictors (2, 3, 5, 7, 8) determined via jackknife methodology.

Summary

Predictive discriminant analysis (PDA) involves a collection of techniques used in multivariate classification, i.e., in predicting membership in well-defined groups for units on which multiple measures are available. The validation (or better yet, assessment) of group membership predictions pertains to two problems: (i) estimating true proportions of correct classifications (i.e., hit rates); and (ii) determining, on the basis of one data set, a good classification rule to use with other data sets.

As in estimating the true validity coefficient in multiple regression/prediction, there are two general approaches to the hit rate estimation problem. The formula approach has been developed only for the two-group case. The external analysis approach may be used in two- or k-group situations. With the latter approach, a classification rule is determined on the basis of one set of data, with hit rate estimates found by applying the rule to different data. Two methods of carrying out an external analysis are suggested: the holdout method, and the leave-one-out (L-O-O) method.

Only in the situation where a very large data set is available should the holdout method be considered. All three computer packages--BMD, SPSS, and SAS--can be used with holdout, or test, samples. If the holdout method is to be used, it is suggested that the test sample be 25-35% of the total sample. In general, the L-O-O method is preferred--only the BMD package employs the L-O-O method.

The form of a classification rule considered in this paper is a set of k linear classification functions (LCFs). The problem of determining a good classification rule has two dimensions: (i) determining good LCF weight estimates; and (ii) determining which predictors to delete from consideration in the final formulation of the rule. Analysis-wise, these two dimensions are considered sort of simultaneously. Jackknife methodology is used to obtain LCF weight estimates, and to assess the variability of the estimates. This assessment, in turn, leads to approximate t statistics which may be used to determine which of the predictors, if any, to delete.

The use of jackknife methodology in determining a good classification rule appears to have some promise. This promise was illustrated herein by the performance of jackknifed LCFs on new data. Of course, more empirical study is called for--considerations that need to be made are magnitude of the true hit rate, ratio of sample size to number of predictors, and differential predictor contribution to classification accuracy.

The computational aspects of hit rate estimation, jackknifing LCFs, assessing LCF weight variability, and applying LCFs to new data are not too difficult. Some results are obtainable via computer package programs. All of the computations needed for this paper were accomplished by a set of two FORTRAN programs written by the current authors. The first program yields the following: (i) internal classification results; (ii) L-O-O classification results; and (iii) classification results based on jackknifed LCFs. Supplementary output includes D-squared values, LCF weights and constants, jackknifed weights and constants, estimated standard errors of the jackknifed weights, and t-ratios for the jackknifed weights. The first program outputs the prior probabilities, internal LCF weights and constants, and jackknifed weights and constants to tape, disk, or cards to be reused in

the second program. The second program uses the two sets of weights and constants to classify new cases, and outputs two sets of classification results. Computational accuracy has been verified; no claim, however, is made for computational or computer efficiency. It should be noted that the programs are somewhat IBM specific and would require modifications for use on other systems.

Appendix

Obtaining a Holdout Analysis

SPSS's DISCRIMINANT

Within DISCRIMINANT the subcommand SELECT is used to specify a variable, the value of which determines whether a unit will be in the analysis sample used to determine the set of k linear classification functions (LCFs), or be in the test sample of units subsequently classified using those LCFs. An example of the subcommand is:

SELECT = VAR(1)/

If such a statement is included in the DISCRIMINANT procedure, those units where the value of the variable VAR is 1 would be used to determine the LCFs, and all other units would be included in the test sample.

Any predefined variable could be used for this purpose. For example, if the variable SEX, with values of 1 for male and 2 for female, was of interest, the user could specify:

SELECT = SEX(1)/

The LCFs would be based on only the male units and the females would then be classified using those LCFs. Another example might involve the variable YEAR. Of course, this is not the usual approach taken for the problem on hand.

A test sample is often a random sample of a specified percent of the total available sample. The problem is to create a variable for which a certain percent of the units take on one value, and every unit has the same chance of being assigned that value. As an example, suppose one wants to base calculation of LCF weights on 70% of the total sample, and use the remaining 30% of the units for estimating true hit rates. The following COMPUTE statement could be used:

```
COMPUTE SET = TRUNC (UNIFORM) (1.43)).
```

[Note: $1.43 \doteq 1/.60$.] This statement will create a variable called SET with possible values of 0 and 1. It is created in a two-step process. First, the UNIFORM function will generate random numbers between 0 and 1.43 and assign one to each unit. [Of the numbers so generated, 70% can be expected to be between 0 and 1.] Second, the TRUNC function will truncate these values to integers. A random 70% (approximately) of the units will take on a value for SET equal to 0. A specification is then made in the DISCRIMINANT procedure,

```
SELECT = SET (0)/,
```

resulting in a 30% test sample. An alternative set of statements, although perhaps less efficient, more clearly illustrates the idea:

```
COMPUTE SET = UNIFORM (100)
IF      (SET LE 70) SET = 0
```

The variable SET could then be used as before in the DISCRIMINANT procedure, again resulting in a 30% test sample. Either method can be altered to obtain different percent test samples. These methods do not yield a test sample of exactly the desired percent, and repeated runs will, of course, produce different results.

BMD's P7M

To generate a test sample with the BMDP7M program, a TRANSFORM paragraph is used (for k = 3 groups and a 30% test sample):

```
/TRANSFORM      PROB = RNDU(783217).
                NEWGROUP = GROUP +3
                TRUE = PROB LE 0.3.
                GROUP = NEWGROUP  IF  TRUE.
```

The process is similar to that in SPSS. First a variable called PROB is created which takes on values between 0 and 1. Then, a new grouping variable NEWGROUP is created. NEWGROUP takes on the value of 4 if GROUP,

the original grouping variable, is equal to 4, 5 if GROUP equals 2, and 6 if GROUP equals 3. In this example, for 30% of the units the GROUP value is reassigned the NEWGROUP value, thus creating three new groups. GROUP now takes on the values 1 through 6. In the GROUP paragraph the statement

USE = 1 to 3.

indicates that only the first three groups will be used as the analysis sample in determining the LCFs, which, in turn, are used to classify the units in the last three groups (i.e., the test sample).

SAS's DISCRIM

For the holdout method with the DISCRIM procedure, the user must externally generate the units to comprise the analysis sample (as well as the test sample). When PROC DISCRIM is run on the analysis data, a statement such as

OUT = CLASTAT;

must be included to create a "data set" where the classification statistics are stored. Then, to apply these statistics to the test sample, the user must specify

DATA = CLASTAT;

and indicate the location of the test data using a statement such as

TESTDATA = NEW;.

the original grouping variable, is equal to 1, 5 if GROUP equals 2, and 6 if GROUP equals 3. In this example, for 30% of the units the GROUP value is reassigned the NEWGROUP value, thus creating three new groups. GROUP now takes on the values 1 through 6. In the GROUP paragraph the statement

```
USE = 1 to 3.
```

indicates that only the first three groups will be used as the analysis sample in determining the LCFs, which, in turn, are used to classify the units in the last three groups (i.e., the test sample).

SAS's DISCRIM

For the holdout method with the DISCRIM procedure, the user must externally generate the units to comprise the analysis sample (as well as the test sample). When DISCRIM is run on the analysis data, a procedure statement such as

```
PROC DISCRIM OUT = CLASTAT;
```

must be used to create a "data set" where the classification statistics are stored. Then, to apply these statistics to the test sample and to indicate the location of the test data, the user must specify

```
PROC DISCRIM DATA = CLASTAT TESTDATA = NEW;
```

References

Afifi, A. A., & Azen, S. P. Statistical analysis. New York: Academic Press, 1979.

Campbell, N. A. Shrunken estimators in discriminant analysis and canonical variate analysis. Applied Statistics, 1980, 29, 5-14.

DiPillo, P. J. The application of bias to discriminant analysis. Communications in Statistics, 1976, A5, 843-854.

DiPillo, P. J. Further applications of bias to discriminant analysis. Communications in Statistics, 1977, A6, 933-943.

Foley, D. H. Considerations of sample and feature size. IEEE Transactions on Information Theory, 1972, 18, 618-626.

Frank, R. E., Massy, W. F., & Morrison, D. G. Bias in multiple discriminant analysis. Journal of Marketing Research, 1965, 2, 250-258.

Habbema, J. D. F., & Hermans, J. Selection of variables in discriminant analysis by F-statistic and error rate. Technometrics, 1977, 19, 487-493.

Hecker, R., & Wegener, H. The validation of classification rates in stepwise discriminant analysis. Biometrical Journal, 1978, 20, 713-727.

Huberty, C. J Discriminant analysis. Review of Educational Research, 1975, 45, 543-598.

Huberty, C. J Issues in the interpretation of discriminant analysis. AERA, 1982.

Huberty, C. J, & Mourad, S. A. Estimation in multiple correlation/prediction. Educational and Psychological Measurement, 1980, 40, 101-112.

Lachenbruch, P. A. An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. Biometrics, 1967, 23, 639-645.

Lachenbruch, P. A., & Mickey, M. R. Estimation of error rates in discriminant analysis. Technometrics, 1968, 10, 1-11.

McKay, R. J., & Campbell, N. A. Variable selection techniques in discriminant analysis II. Allocation. British Journal of Mathematical and Statistical Psychology, 1982, 35, 30-41.

McLachlan, G. J. Confidence intervals for the conditional probability of misallocation in discriminant analysis. Biometrics, 1975, 31, 161-167.

McLachlan, G. J. On the relationship between the F test and overall error rate for variable selection in the two-group discriminant analysis. Biometrics, 1980, 36, 501-510.

Michaelis, J. Simulation experiments with multiple group linear and quadratic discriminant analysis. In T. Cacoullos (Ed.), Discriminant analysis and applications. New York: Academic Press, 1973.

Miller, R. G. The jackknife-a review. Biometrics, 1974, 61, 1-15.

Mosteller, F., & Tukey, J. W. Data analysis and regression. Reading, Mass.: Addison-Wesley, 1977.

Murray, G. D. A cautionary note on selection of variables in discriminant analysis. Applied Statistics, 1977, 26, 246-250.

Schaafsma, W., & van Vark, G. N. Classification and discrimination problems with applications, Part IIa. Statistica Neerlandica, 1979, 33, 91-126.

Stone, M. Cross-validating choice and assessment of statistical predictions. Journal of the Royal Statistical Society, Ser. B., 1974, 36, 111-133.

Wildt, A. R., Lambert, Z. V., & Durand, R. M. Applying the jackknife statistic in testing and interpreting canonical weights, loadings, and cross-loadings. Journal of Marketing Research, 1982, 19, 99-107.

Footnote

[1]In this paper, only linear rules are considered; i.e., rules based on the assumption of equal group covariance matrices--see Huberty (1975) for a discussion of quadratic classification rules. LCF formulation is given in the next major section of this paper.

Table 1

Hit Rates for the Two-Group Situation

	$\hat{P}_1$	$\hat{P}_2$	$\hat{P}_T$
Internal	.88	.30	.66
Formula [1]	.85	.37	.67
Formula [2']	.53	.45	.47
Holdout[a]	.78	.00	.58
L-O-O	.79	.20	.57
Expected	.39	.14	.53

[a]30% test sample.

Table 2

Hit Rates for the Three-Group Situation

	$\hat{P}_1$	$\hat{P}_2$	$\hat{P}_3$	$\hat{P}_T$
Internal	.82	.10	.50	.53
Holdout[a]	.44	.00	.50	.26
L-O-O	.76	.00	.38	.44
Expected	.18	.07	.10	.35

[a]30% test sample.

Table 3

LCF Estimation

	Predictor	Internal	Jackknife
LCF_1	1	-.06	-.05
	2	.10	.08
	3	.95	.76
	4	.26	.30
	5	.96	.69
	6	-.63	-.43
	7	.14	.11
	8	2.59	1.71
	Constant	-14.00	-10.69
LCF_2	1	-.01	-.02
	2	.08	.06
	3	.87	.69
	4	.31	.35
	5	.82	.63
	6	-.56	-.39
	7	.10	.08
	8	3.19	2.20
	Constant	-13.54	-10.33
LCF_3	1	.00	.00
	2	.05	.03
	3	.96	.77
	4	.26	.27
	5	.76	.57
	6	-.52	-.40
	7	.10	.07
	8	3.07	2.18
	Constant	-14.13	-10.89

Table 4

t-Ratios for Jackknifed Weights

(Original Predictor Set)

Predictor	LCF 1	2	3
1	.66	.20	.02
2	.94	.76	.40
3	2.05	1.68	1.96
4	.70	.80	.64
5	1.04	1.10	.96
6	.75	.77	.83
7	1.12	.92	.77
8	.73	.95	.93

Table 5

Jackknifed Weights and t-Ratios

(Predictor Subset)

	Predictor	Weights	t-Ratios
LCF_1	2	.08	1.17
	3	.76	2.18
	5	.67	1.34
	7	.09	.98
	8	1.16	.55
LCF_2	2	.08	1.41
	3	.72	1.91
	5	.64	1.42
	7	.05	.58
	8	1.83	.93
LCF_3	2	.07	1.00
	3	.80	2.16
	5	.58	1.13
	7	.04	.42
	8	1.79	.86